CLUSTERING AND SYNCHRONIZING MULTI-CAMERA VIDEO VIA LANDMARK CROSS-CORRELATION

Nicholas J. Bryan[1*] Paris Smaragdis[2,3] Gautham J. Mysore[3]

[1] Center for Computer Research in Music and Acoustics, Stanford University
[2] University of Illinois at Urbana-Champaign
[3] Advanced Technology Labs, Adobe Systems Inc.

Introduction

Problem: Cluster and synchronize multiple videos of the same event as fast and efficiently as possible.

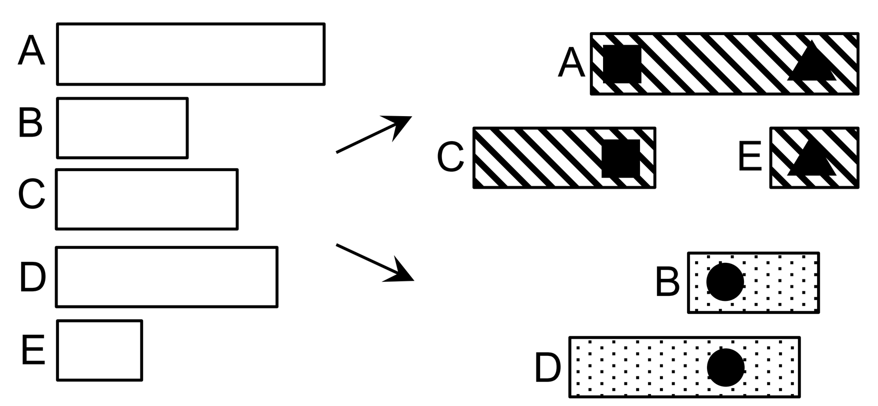


Fig. 1. Clustering and synchronizing an unograized video/audio collection.

Proposed Method

Method: Use landmark-based audio fingerprinting for a fast and efficient method to both cluster and synchronize videos.

More specifically, the method is outlined in five sections:

- Non-linear transform
- Time-difference-of-arrival estimation
- Agglomerative clustering
- Efficient computation
- Synchronization refinement

Non-Linear Transform

The audio signal is transformed into a set of combinatorially-paired frequency peak onsets [4] and stored in a sparse high-dimensional discrete time signal $\mathbf{L}(t)$.

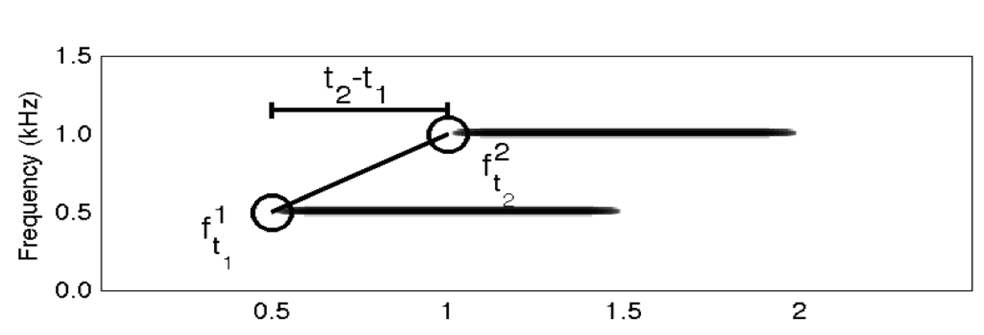
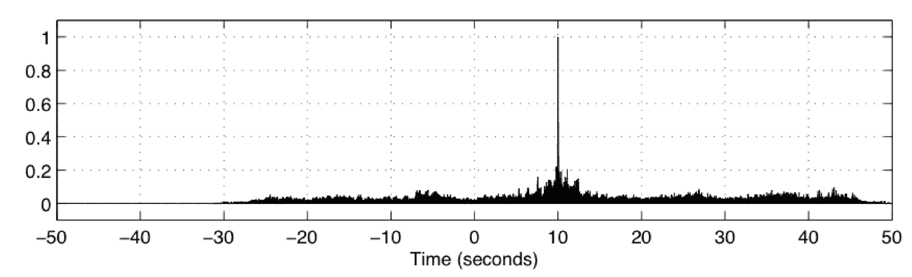
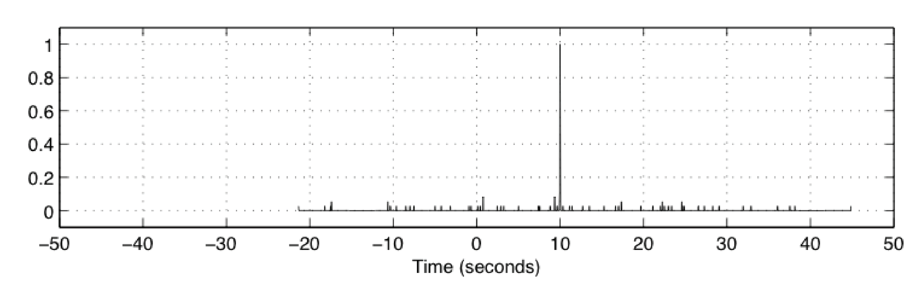


Fig. 2. An example landmark created from a 1kHz tone starting at .5 seconds and a 2kHz tone starting at 1 second.

Time-Difference-Of-Arrival Estimation



(a) Normalized absolute time-domain cross-correlation.



(b) Normalized landmark cross-correlation.

Fig. 3. Example cross-correction of speech signals.

$$R_{\mathbf{L}_i,\mathbf{L}_j}(t) = \sum_{\tau=-\infty}^{\infty} \mathbf{L}_i(\tau)^{\mathrm{T}}\mathbf{L}_j(t+\tau)$$

$$\hat{t}_{ij} = \arg\max_{t} \; R_{\mathbf{L}_i,\mathbf{L}_j}(t)$$

Agglomerative Clustering

Initialize each recordings as a separate cluster and then merged into successively larger clusters if $\hat{R}_{\mathbf{L}_i,\mathbf{L}_j} \geq \theta$, where $\hat{R}_{\mathbf{L}_i,\mathbf{L}_j} = \max_t \; R_{\mathbf{L}_i,\mathbf{L}_j}(t)$.

Other useful decision rules include:

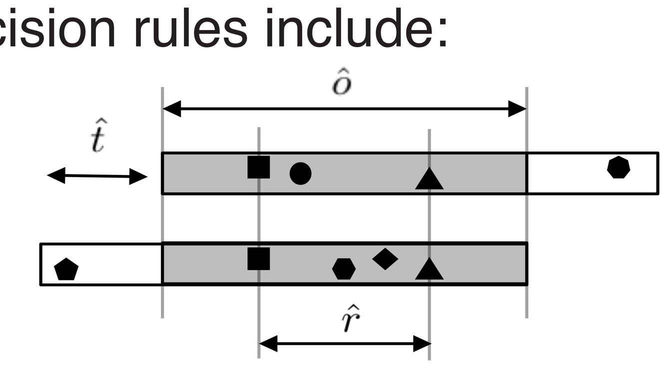


Fig. 4. Two different recordings of the same event.

1. Reject merges with a small percentage of total matching landmarks (in both files) in the overlap region $\hat{o}$.
2. Reject merges with a small overall time range $\hat{r}$ defined by the set of matching landmarks.
3. Reject merges with a small overlap region $\hat{o}$.

Efficient Computation

- Cross-correlation traditionally requires $O(N^2)$ operations or O(N log N) operations for FFT-based correlation, where N is the file length.
- Slight modification to the map structure of [4] can be interpreted as a sparse correlation method, requiring only the initial cost of building the map structure + O(N) operations, where N is the number of matching landmarks.

Synchronization Refinement

Refinement is required for clusters of three or more when:

- pairwise TDOA estimates do not satisfy all triangle equalities
- one or more TDOA estimates within any cluster is unknown.

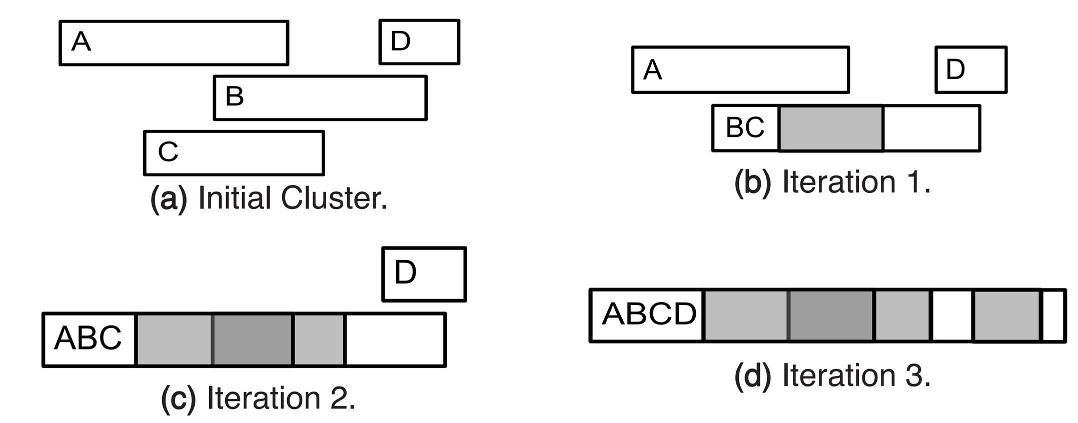


Fig. 5 Match-and-merge synchronization refinement.

Match-and-Merge Algorithm

1. Find the most confident TDOA estimate $\hat{t}_{ij}$ within the cluster in terms of $\hat{R}_{\mathbf{L}_i,\mathbf{L}_j}$ or similar confidence score.
2. Merge the landmark signals $\mathbf{L}_i(t)$ and $\mathbf{L}_j(t)$. First time shift $\mathbf{L}_j(t)$ by $\hat{t}_{ij}$ and then multiply or add the two signals together (depending on the desired effect).
3. Update the remaining TDOA estimates and confidence scores to respect the file merge.
4. Repeat until all files within the cluster are merged.

Evaluation

Speech dataset - 180 speech recordings from a film set with 114 clusters averaging 20-40 seconds in length.

Music dataset - 23 cell-phone recordings of three live music concerts averaging 3-6 minutes in length.

	Speech	Music	Speech + Music
Precision	100.0 %	100.0 %	100.0 %
Recall	97.0 %	100.0 %	99.2 %
F-Score	98.5 %	100.0 %	99.6 %

Table 1. Precision, recall, and F1-scores.

	Speech	Music	Speech + Music
Proposed	47.0 / 164.6	41.1 / 146.5	90.1 / 152.7
Traditional	1550 / 5.0	197 / 30.5	3600 / 3.9

Table 2. Computation time (s) and throughput (s/s).

Conclusions

- Cluster and synchronize audio via landmark audio fingerprinting.
- Improvements on event identification and synchronization refinement.
- Presented within the framework of cross-correlation.
- High accuracy with efficient computation.

References

[1] P. Shrestha, M. Barbieri, and H. Weda, "Synchronization of multi-camera video recordings based on audio," in *Proc. 15th Intl. Conf. on Multimedia*, 2007.

[2] J. Haitsma and T. Kalker, "A highly robust audio fingerprinting system with an efficient search strategy," *Journal of New Music Research*, vol. 32, no. 2, 2003.

[3] L. Kennedy and M. Naaman, "Less talk, more rock: automated organization of community-contributed collections of concert videos," in *Proc. 18th Conf. on World Wide Web*, 2009.

[4] A. L. Wang, "An Industrial-strength audio search algorithm," in *Proc. 4th Int. Symposium on Music Information Retrieval (ISMIR)*, October 2003.

[5] C.V. Cotton and D.P.W. Ellis, "Audio fingerprinting to identify multiple videos of an event," in *Proc. ICASSP*, March 2010.

[6] P. Cano, E. Batlle, T. Kalker, and J. Haitsma, "A review of audio fingerprinting," *J. VLSI Signal Processing Sys*, vol. 41, November 2005.

*This work was performed while interning at Adobe Systems Inc.